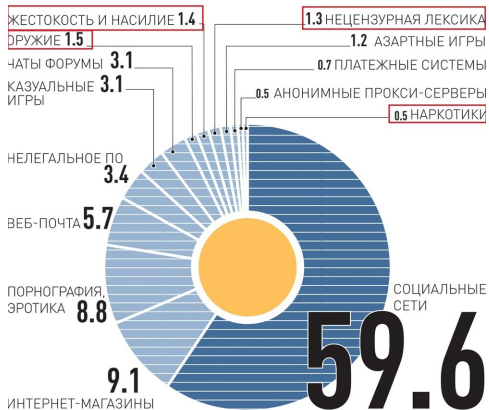


Классификация текстового контента

Александр Смирнов и Федор Жилкин

24.05.2019г

Введение



Цели

- ▶ Ограничить детей от взрослого текстового контента в интернете
- ▶ Получить опыт
 - ▶ Бинарная классификация текста
 - ▶ Сбор данных для обучения
 - ▶ Написание Python-библиотеки
 - ▶ Написание расширения для Chrome
 - ▶ Написание Python-сервера для приёма запросов

Сравнение с аналогами

- ▶ Ограничения на поиск
 - ▶ Семейный поиск Яндекс
 - ▶ Безопасный поиск Google
- ▶ Контентная фильтрация
 - ▶ Traffic Inspector
 - ▶ Интернет Цензор

Сравнение с аналогами (2)

The screenshot shows a software interface for network traffic analysis. On the left is a tree view of the interface, and on the right is a table of protocols and ports.

Имя	Протокол	Порт назнач.	Порт источн.	Примечание
IMAP client (POP/50)	UDP	50		
FTP client (TCP/21)	TCP	21	21	
FTP server (TCP/21)	TCP			
FTP-DATA client (TCP/20)	TCP	20		
FTP-DATA server (TCP/20)	TCP		20	
HTTP client (TCP/80)	TCP	80		
HTTP server (TCP/80)	TCP		80	
HTTPS client (TCP/443)	TCP	443		
HTTPS server (TCP/443)	TCP		443	
ICMP	ICMP			
ICQ/Authorization (TCP/5190-5191)	TCP	5190-5191		
ISMP	Заданный порт IP (2)			
IMAP client (TCP/143)	TCP	143		
IMAP server (TCP/143)	TCP		143	
MailBox name and datagram service client (UDP/119)	UDP	117-119		
MailBox session service client (TCP/117-119)	TCP	117-119		
POP3 client (TCP/110)	TCP		110	
POP3 server (TCP/110)	TCP		110	
Remote Administrator client (TCP/4899)	TCP	4899		
SMTP client (TCP/25)	TCP	25	25	
SMTP server (TCP/25)	TCP		25	
SSH client (TCP/22)	TCP	22		
Telnet server (TCP/23)	TCP		23	
Windows Remote Desktop client (RDP) (TCP/3389)	TCP	3389		
Windows Remote Desktop server (RDP) (TCP/3389)	TCP		3389	

Рис.: Пример интерфейса схожей программы

Задачи

- ▶ Провести анализ возможных решений для классификации текста
- ▶ Собрать рассказы для взрослых и обычные рассказы
- ▶ Написать Python-сервер, использующий обученную модель для ответа на запросы от расширения
- ▶ Сделать расширение для Chrome, обращающееся к серверу

Сбор данных

- ▶ Рассказы для взрослых берём с сайта ideer.ru
- ▶ Рассказы для широкого круга читателей берём с множества сайтов по разным тематикам

Анализ подходов к классификации текста

- ▶ Rule-based
- ▶ Machine Learning based
- ▶ Hybrid systems

Характеристики сравнения эффективности

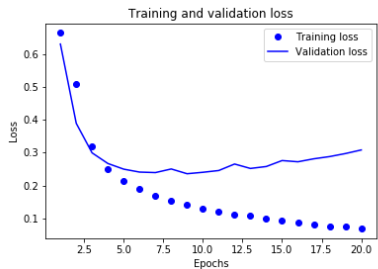
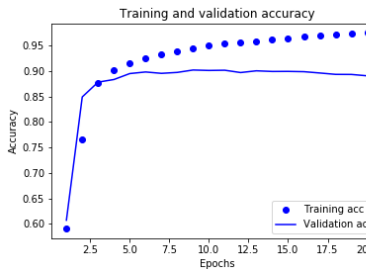
- ▶ Accuracy – общая точность классификатора
- ▶ Recall – отношение заблокированных взрослых сайтов к общему количеству взрослых сайтов (% классифицированных взрослых сайтов)
- ▶ Precision – отношение заблокированных взрослых сайтов к числу всех заблокированных сайтов (точность блокировки)
- ▶ F1 Score – среднее гармоническое между Precision и Recall, для учёта и того, и другого в одной величине

Сравнение

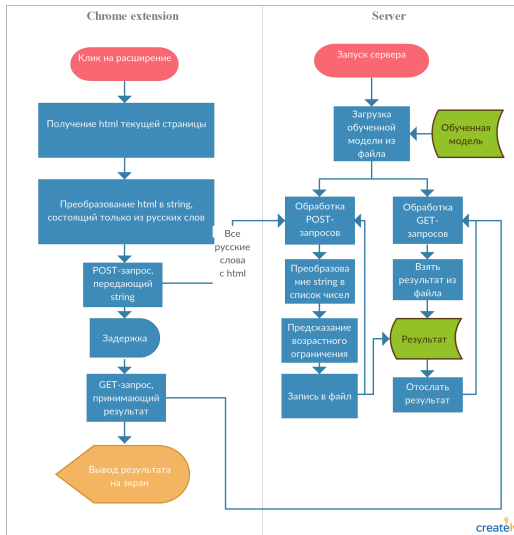
- ▶ Random model – случайный выбор блокировать/не блокировать
- ▶ Rule-based model – блокировка по списку непотребных слов
- ▶ Classifier – 3-х слойная обычная сеть
- ▶ Upgraded Classifier – Classifier, из словаря которой были исключены самые частые слова и добавлена ненормативная лексика

	F1 Score	Accuracy	Recall	Precision
Random model	0.58	0.51	0.58	0.58
Rule-based model	0.06	0.41	0.03	1.0
Classifier	0.90	0.88	0.93	0.87
Upgraded Classifier	0.91	0.90	0.92	0.91

Результаты обучения



Взаимодействие расширения и сервера



Итоги

- ▶ Федор
 - ▶ Сбор данных
 - ▶ Библиотека
- ▶ Александр
 - ▶ Классификация текста
 - ▶ Сервер
 - ▶ Расширение

Результаты

- ▶ Сделано расширение для Chrome – <https://github.com/SmirnovAlexander/PoemClassifier>
- ▶ Сделана библиотека на pyPi – <https://pypi.org/project/TalesParse/>
- ▶ Собраны рассказы на kaggle – <https://www.kaggle.com/idoldev/adult-and-child-russian-tales-dataset-with-label>
- ▶ Написан сборщик рассказов – https://github.com/Feodoros/Scraping_Tales